International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN (P): 2249-6831; ISSN (E): 2249-7943
Vol. 7, Issue 4, Aug 2017, 67-70

TRANS STELLAR
•Journal Publications •Research Consultancy

REVIEW PAPER ON DECISION TREE DATA MINING ALGORITHMS TO IMPROVE ACCURACY IN IDENTIFYING CLASSIFIED INSTANCES USING LARGE DATASET

GURPREET SINGH[1] & ER. RAJWINDERKAUR[2]

[1]*Professor, Department of Computer Science & Engineering, St. Soldier Institute of Engineering & Technology, Jalandhar Punjab, India*

[2]*M.Tech Scholar, Department of Computer Science & Engineering, St. Soldier Institute of Engineering & Technology, Jalandhar, Punjab, India*

ABSTRACT

The CART distance based algorithm with the classification tree paradigm based on the C45 algorithm. The CART algorithm is used as a preprocessing algorithm in order to obtain a modified training database for the posterior learning of the classification tree structure. Then the incorrectly classified instances are duplicated with the previous data set and finally C45 is applied to complete the classification procedure of biomedical data.

KEYWORDS: The Hierarchal Model of Decisions . Data Mining is a Technology that Draws Out Information from Colossal Amount of Gigantic Data and Remolds it into a Human Understandable Form

Received: Jul 08, 2017; **Accepted:** Jul 25 2017; **Published:** Aug 01, 2017; **Paper Id.:** IJCSEITRAUG20179

Original Article

INTRODUCTION

Decision trees basically us the hierarchal model of decisions and their consequences. The structure of decision tree includes branch, root node and leaf node. Attributes test is denoted on each interval node, the test outcome is denoted by branch and class labels are shown by leaf node. The topmost node is the root node of the tree. The tree learning is done by dividing the source into set which are generally based on a test of attribute value. Data mining is a technology that draws out information from colossal amount of gigantic data and remolds it into a human understandable form. There are many other terminologies identical to data mining-knowledge mining from data, knowledge extraction.

Review on

- Enhanced decision tree algorithm which will work on large scale high. An algorithm can be made with certain split selection methods involved from the literature which includes algorithms like C4.5 and CART.
- Enhance the efficiency with a new classifier that combines the CART distance based algorithm with the classification tree paradigm based on the C45 algorithm.
- Reducing present sum of square error- the proposed algorithm gives reduced sum of square error as compare to the CART and C4.5 classification algorithm which means that the new algorithm gives more accuracy.
- Enhancement in the efficiency of decision tree construction- various pruning techniques are proposed which can help in the improvement of decision tree construction.
- Research methodology is the organized way to solve a research problem. It is a conceptual way which tells

that how the research is done by the researcher**.**

C4.5 ALGORITHM

C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a "split information" value defined analogously with Info (D) as

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right).$$

This value represents the potential information generated by splitting the training data set, D, into v partitions, corresponding to the v outcomes of a test on attribute A. Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D. It differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning. The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}.$$

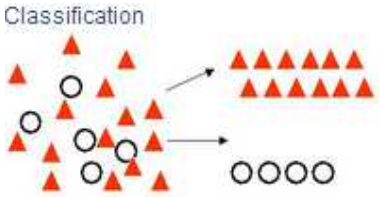
The Gini index is used in CART. Using the notation described above, the Gini index measures the impurity of D, a data partition or set of training tuples, as

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,$$

Classification & Regression Trees (CART)

CART were invented independently of one another at around the same time, yet follow a similar approach for learning decision trees from training tuples. These two cornerstone algorithms spawned a flurry of work on decision tree induction.

Decision tree induction can be adapted so as to predict continuous (ordered) values, rather than class labels. There are two main types of trees for prediction—regression trees and model trees. Regression trees were proposed as a component of the CART learning system. (Recall that the acronym CART stands for Classification and Regression Trees.) Each regression tree leaf stores a continuous-valued prediction, which is actually the average value of the predicted attribute for the training tuples that reach the leaf. Since the terms "regression" and "numeric prediction" are used synonymously in statistics, the resulting trees were called "regression trees," even though they did not use any regression equations. By contrast, in model trees, each leaf holds a regression model—a multivariate linear equation for the predicted attribute. Regression and model trees tend to be more accurate than linear regression when the data are not represented well by a simple linear model.




```
Size of the tree :      12


Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          84               85.7143 %
Incorrectly Classified Instances        14               14.2857 %
Kappa statistic                          0.5393
Mean absolute error                      0.2339
Root mean squared error                  0.342
Relative absolute error                 61.1713 %
Root relative squared error             78.4536 %
Total Number of Instances               98

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0.56       0.839     1         0.913    good
  0.44      0          1         0.44      0.611    bad

=== Confusion Matrix ===

  a  b   <-- classified as
 73  0 |  a = good
 14 11 |  b = bad
```

Figure 1: Results of J45 Algorithm in Weka

```
Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          79               80.6122 %
Incorrectly Classified Instances        19               19.3878 %
Kappa statistic                          0.4381
Mean absolute error                      0.3038
Root mean squared error                  0.3898
Relative absolute error                 79.4424 %
Root relative squared error             89.4057 %
Total Number of Instances               98

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.918     0.52       0.838     0.918     0.876    good
  0.48      0.082      0.667     0.48      0.558    bad

=== Confusion Matrix ===

  a  b   <-- classified as
 67  6 |  a = good
 13 12 |  b = bad
```

Figure 2: Results of Decision Stump Algorithm in Weka

CONCLUSIONS

In this comparative study found that J45 gives the better performance as compare to Decision Stump with minimum error rate or high accuracy, maximum percentage of correctly classified instances on same data set and parameters.

REFERENCES

1. *U. Fayyad, G. Piatetsky-Shapiro, and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence. All rights reserved. 0738-4602-1996*
2. *S. Lallich, O. Teytaud, Évaluation et validation de l'intérêt des règlesd'as sociation*
3. *Osada, R., Funkhouser, T., Chazelle, B. et Dobkin, D. ((Matching 3D Models with Shape Distributions)). Dans Proceedings of the International Conference on Shape Modeling & Applications (SMI '01), pages 154–168. IEEE Computer Society, Washington, DC, Etat-Unis. 2001.*
4. *W.Y. Kim et Y.S. Kim. A region-based shape descriptor using Zernike moments. Signal Processing : Image Communication, 16 :95–100, 2000.*
5. *Khotanzadet Y.H. Hong. Invariant image recognition by Zernike moments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(5), May 90.*
6. *N Pasquier, Y Bastide, R Taouil, L Lakhal - Database Theory ICDT'99, 1999 – Springer*
7. *Ji Dan, Qiu JianlinetGu Xiang, Chen Li, He Peng. (2010) A Synthesized Data Mining Algorithm based on Clustering and Decision tree. 10th IEEE International Conference on Computer and Information Technology (CIT 2010)*
8. *Timothy C. Havens et James C. Bezdek. Fuzzy c-Means Algorithms for Very Large Data*